Bayesian partial factor regression

P. Richard Hahn
Department of Statistical Science, Duke University

Carlos Carvalho
McCombs School of Business, The University of Texas, Austin

Sayan Mukherjee
Departments of Statistical Science,
Computer Science, Mathematics, and
Institute for Genome Sciences & Policy, Duke University

Abstract

A Bayesian linear regression model is developed that cleanly addresses a long-recognized and fundamental difficulty of factor analytic regression – the response variable could be closely associated with the least important principal component. The model possesses inherent robustness to the choice of the number of factors and provides a natural framework for variable selection of highly correlated predictors in high dimensional problems. In terms of out-of-sample prediction, the model is demonstrated to be competitive with partial least squares, ridge regression, and standard factor models under data regimes for which each of those methods excels; thus representing a promising default regression tool. By incorporating point-mass priors on key parameters this model permits variable selection in the presence of highly correlated predictors, as well as estimation of the sufficient dimension, in the $p \gg n$ setting.

1 The Predictor Distribution's Role in Linear Regression

1.1 Introduction

In prediction problems with more predictors than observations, it can sometimes be helpful to use a joint model, $f(Y, X)$, rather than a purely conditional model, $f(Y \mid X)$. This approach is motivated by the fact that that in many situations the marginal predictor distribution $f(X)$ can provide useful information about the parameter values governing the conditional regression [Liang et al., 2007]. However, this marginal distribution can sometimes provide misinformation which can lead conditional inferences astray. Here, we explore these ideas in the context of linear factor models, to understand in a well-studied model how these ideas play out. The end result is a model-based analogue of the shrinkage principal component model of George and Oman [1996]. The resulting Bayesian model, which we call partial factor regression, performs well across a wide range of covariance structures, outperforming standard factor models and ridge regression.

Suppose we wish to perform a linear regression and that we have more predictors, p, than we have observations, n. Further assume that the errors are Gaussian, with residual variance σ^2. The sampling model may then be written as

$$Y_i \quad = \quad X_i\beta + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2). \tag{1}$$

This model is a pure regression model in the sense that the sampling model is written conditionally on observed predictors X. For this reason, the $p \times n$ matrix of observed predictors – for a sample of size n – is often referred to as a "design matrix"; because the statistical properties of the model do not change

with respect to specific values of this matrix, the analyst is hypothetically at liberty to choose particular values. Of course, in many regression applications we do not have such freedom, so that what we deal with are in actuality stochastic predictors. This fact raises the question of whether or not we should take this stochasticity into account when selecting our statistical model. What role, if any, should the marginal distribution of the predictors play in a linear regression for prediction? More generally, which aspects of the covariance structure of the predictors are relevant for predicting the response?

The first step towards incorporating the marginal distribution of the predictors into our regression model is to understand the role played by the design matrix in the pure regression case. In a Bayesian framework our sampling model is complemented with a prior $\pi(\beta, \sigma^2)$. Here we study the common choice of a conjugate Normal-Inverse-Gamma prior:

$$\beta \mid \sigma^2, S_0 \sim \mathrm{N}(0, \sigma^2 S_0) \tag{2}$$

$$\sigma^2 \sim \mathrm{IG}(a, b). \tag{3}$$

From a predictive perspective, the role of the prior distribution on β is to provide regularization. One common choice takes the prior covariance matrix to be $S_0 = \tau^{-1} I$, where τ can additionally be assigned a hyperprior. The posterior mean in this case gives the well-known ridge estimator when the prior mean is zero with:

$$\tilde{\beta} = \mathrm{E}(\beta \mid Y, X) = (XX^t + \tau^{-1} I)^{-1} XX^t \hat{\beta}, \tag{4}$$

where $\hat{\beta}$ is the least-squares estimator

$$\hat{\beta} = (XX^t)^{-1} XY.$$

Another popular choice of S_0 gives Zellner's g-prior [Zellner, 1986] by setting

$$S_0 = (XX^t)^{-1}/g \tag{5}$$

yielding the estimate

$$\bar{\beta} = \mathrm{E}(\beta \mid Y, X) = (1+g)^{-1} \hat{\beta}. \tag{6}$$

This prior cannot be used unmodified when $p > n$ due to the singularity of $(XX^t)^{-1}$ in which case pseudoinverses may be used [Liang et al., 2008].

It is straightforward to show [Hastie et al., 2001] that the ridge estimator downweights the contribution of the directions in (observed) predictor space with lower sample variance. The rational for this behavior is described as follows:

> Ridge regression protects against the potentially high variance of gradients estimated in the short directions. The implicit assumption is that the response will tend to vary most in the directions of high variance in the inputs.

The g-prior, by contrast, shrinks β more in directions of high sample variance in the predictor space a priori, which has the net effect of shrinking the orthogonal directions of the design space equally regardless of whether the directions are long or short. This reflects the belief that higher variance directions in predictor space need not influence the response variable more than the directions of lower variance.

The justifications for ridge regression and the g-prior conflate the observed design space with the pattern of stochastic covariation characterizing the random predictor variable. Our goal here is to realize the benefit of regularizing estimates in directions of low sample variance, while not over-regularizing regions of predictor space with weak stochastic covariance structure. Teasing apart these two aspects of the problem, by conditioning on X and $\Sigma_X \equiv \mathrm{cov}(X)$ separately, leads to the partial factor framework.

1.2 Conditioning a regression on Σ_X

For simplicity, we will assume that the response and the predictors follow a joint Normal distribution with covariance Σ and, without loss of generality, mean zero. This setting permits us to think concretely about the dependence structure between our predictors and between the predictors and the response, and

provides the underpinning for more sophisticated models. In particular, the regression, as a function of the covariance, can be expressed as:

$$\begin{aligned} \Sigma = \text{cov}\begin{pmatrix} X \\ Y \end{pmatrix} &= \begin{bmatrix} \Sigma_X & V^t \\ V & \omega \end{bmatrix}, \\ E(Y \mid X) &= X\beta, \\ \beta &= \Sigma_X^{-1} V^t. \end{aligned} \tag{7}$$

Indeed, in this setting what we really have is a covariance estimation problem, whereupon the regression comes along for free; the focus of this paper is to describe how to estimate this covariance in a model-based way that makes sure not to neglect the implied regression in the process. The usual least-squares estimator in this context mirrors the form of this conditional mean, plugging in estimators for Σ_X and V in the appropriate places:

$$\hat{\beta} = (XX^t)^{-1}XY = \widehat{\Sigma}_X^{-1}\hat{V}^t, \tag{8}$$

where $\hat{\Sigma}_X \equiv n^{-1}XX^t$ and $\hat{V} \equiv n^{-1}XY$.

If Σ_X is in hand, we may reparametrize our problem in terms of decorrelated predictors. Let $\tilde{X} = L^{-t}X$, where $L^tL = \Sigma_X$ is the Cholesky decomposition of the covariance matrix. Then, placing an independent prior on the coefficients of this regression yields something that might reasonably be called a population parameter g-prior on the original regression coefficients:

$$\begin{aligned} \begin{pmatrix} \tilde{X} \\ Y \end{pmatrix} &\sim \text{N}(0, \tilde{\Sigma}), \\ \tilde{\Sigma} &= \begin{bmatrix} I & \alpha^t \\ \alpha & \omega \end{bmatrix}, \\ (\alpha \mid \sigma^2) &\sim \text{N}(0, \sigma^2\tau\text{I}). \end{aligned} \tag{9}$$

This specification implies that

$$(\beta \mid \Sigma_X, \sigma^2) \sim \text{N}(0, \sigma^2\tau\Sigma_X^{-1}). \tag{10}$$

To see the impact of using Σ_X^{-1}, as opposed to $n\hat{\Sigma}_X^{-1}$ used by the g-prior, we can look at the expression for the implied posterior mean (assuming for simplicity that $\tau = \sigma^2 = 1$):

$$\text{E}(\beta \mid Y, X, \Sigma_X) = (I + n\Sigma_X^{-1}\hat{\Sigma}_X)^{-1}(\Sigma_X^{-1}V_0 + n\Sigma_X^{-1}\hat{V}), \tag{11}$$

$$\beta_0 = \Sigma_X^{-1}V_0, \tag{12}$$

where V_0 is chosen a priori and determines the prior mean of the regression coefficients. Because Σ_X and $\hat{\Sigma}_X$ are never identical, we still get shrinkage in different directions, thus combatting the "high variance of gradients estimated in short directions" while not having to assume that any direction in predictor space is more or less important a priori.

Of course, we never actually have Σ_X in hand, so we infer it conditional on the observed data. When Σ_X is high dimensional, the choice of prior is important in providing adequate regularization properties.

1.3 Normal factor models for covariance estimation and regression

A multivariate Normal distribution may be written in *factor form* as:

$$X_i = Bf_i + \nu_i, \quad \nu_i \overset{iid}{\sim} \text{N}(0, \Psi) \tag{13}$$

$$f \sim \text{N}(0, I_k). \tag{14}$$

The matrix B is a $p \times k$ real-valued matrix ($p \gg k$) and Ψ is diagonal. Conditional on B and f, the elements of each observation are independent. Integrating over f, we see

$$\Sigma_X = BB^t + \Psi. \tag{15}$$

When $k = p$ this form is unrestricted so that any positive definite matrix can be written as in (15).

If we further assume that the p-dimensional predictors influence on the response Y only through the k-dimensional latent variable f, we arrive at the Bayesian factor regression model of West [2003]:

$$\begin{aligned} Y_i &= \theta f_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2) \\ \Sigma = \mathrm{cov}\begin{pmatrix} X \\ Y \end{pmatrix} &= \begin{bmatrix} BB^t + \Psi & V^t \\ V & \omega \end{bmatrix}, \\ V &= \theta B^t, \\ \omega &= \sigma^2 + \theta\theta^t. \end{aligned} \tag{16}$$

In that paper it is shown that as $|\Psi| \to 0$ this model yields a singular value regression, where the response is regressed on the top k (left) singular vectors of the design matrix. Here θ is a $1 \times k$ row vector; effectively it is just an additional row in the loadings matrix. Factor models generally and Bayesian factor regression specifically, have been an area of active research for many years [Lopes, 2003], found many useful applications Aguilar and West [2000], Carvalho et al. [2008], and continue to see new developments [Bhattacharya and Dunson, 2010, Fruhwirth-Schnatter and Lopes, 2009].

The assumption that our predictors relate to the response via a k-dimensional parameter means that rather than n observations with which to estimate the p-dimensional regression parameter β, as in the conditional regression case, we have $n \times p$ observations – because each predictor dimension is now part of the likelihood – with which to estimate the $k \times p$ elements of the loadings matrix, B. The intuition is simply this: if spotting dominant trends in predictor space is easy, and our response depends on only a small number of these, then the regression problem should also be easy. Analogous to ridge regression, the assumption is that the dominant directions of variability in predictor space are most associated with the response variable; moreover, this space is restricted to be, at most, k-dimensional.

For the present purposes a toy example illustrates the principles at work.

Example Let $p = 20$, $n = 10$ and $k = 1$. Assume that $Y = X_{16}$. While ten observations are not much with which to learn about a 19 dimensional hyperplane – the regression coefficients β, the fact that there exists a one dimensional subspace characterizing the covariation means that our problem splits into 20 independent simple regressions, each with ten observations.

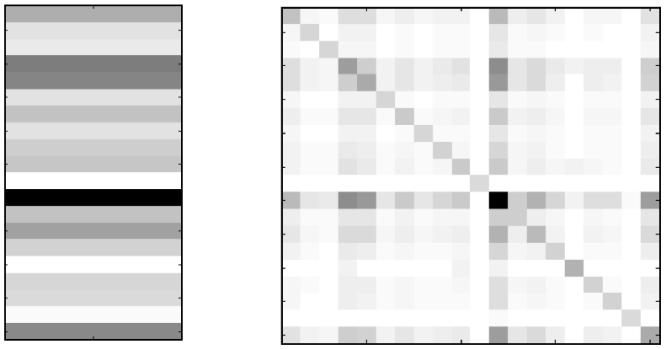

Figure 1: The single factor loading column at left. At right, the implied covariance structure, assuming $\Psi = I$ for illustration purposes.

Because the response need not vary most in these directions, the factor model, seen as a regression, can badly over-regularize as we will see next.

1.4 Over-regularization and the least eigenvalue scenario

Over the years, many statisticians [Hotelling, 1957, Cox, 1968, Jolliffe, 1982] have counseled that "there is no logical reason why the dependent variable should not be closely tied to the least important principal component [Cox, 1968]." We may witness this effect in simple examples.

Example Consider returns on petroleum in the United State and in Europe and assume we are interested in estimating the spread for trading purposes. Let $X = (X_1, X_2)$, where X_1 and X_2 is the price in the U.S. and in Europe, respectively, so that we want to predict $X_1 - X_2$. If we consider the correlation matrix, the first principal component will be given by $X_1 + X_2$ with variance $\frac{1+r}{2}$ while the second component is $X_1 - X_2$ with variance $\frac{1-r}{2}$ where r is the correlation between the two prices. For r near one, a regression based on only the first principal component will discard all the relevant information, because the second principle component is the one of interest [Forzani, 2006].

We see that the bias incurred by throwing away the second principal component is much bigger than the reduction in variance incurred by its elimination. We can observe the same phenomenon in the factor model setting by revisiting the earlier example where the factor model worked well; it is easily modified it so that it fails as a regression model for the response variable.

Example Consider the set up from the first example, but with a modified underlying factor loadings matrix: in addition to the column shown before, we introduce a second column, with only four non-zero entries, as shown. The first column has a (Euclidean) norm near 3.5 while the new second column has norm only near 1. Again, let $Y = X_{16}$. The implied regression coefficients implied by a one versus two factor model show stark discrepancies. Small changes in the joint correlation structure can translate into practically large changes in a univariate conditional regression implied by that joint structure.

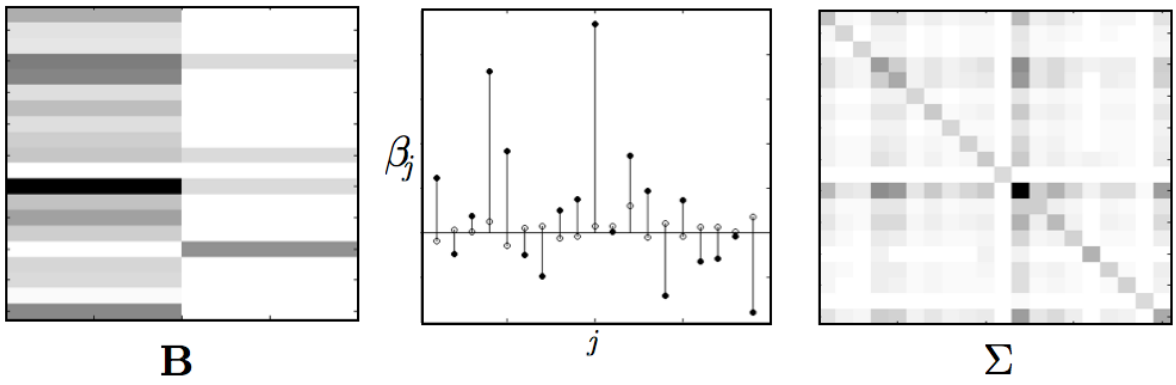

Figure 2: The first panel shows the newly added second factor to the "true" model. The second panel shows the implied regression coefficients for element 16 under the single factor versus double factor model. The single factor coefficients are shown with open circles. The final panel shows that in terms of the overall covariance, the difference of adding the second column is visually minor.

Notice while this model adequately captures the *joint* distribution; the incorrect assumption that $k = 1$ renders it completely inappropriate as a regression model. Note too that this happens even though the signal in the second column of B is quite strong. Setting $m = 1$ when we ought to have set $m = 2$ causes the model to ignore all but the most dominant direction in terms of explaining the joint correlation structure.

This behavior is equally likely to crop up in real data analyses. We conjecture that factor analysis of gene expression studies [Carvalho et al., 2008] were unsuccessful at identifying associations between gene expression and clinical outcomes due in part to the least-eigenvalue problem.

1.5 Choosing k

The problem of over-regularization cannot be overcome simply by employing a factor model with an unknown number of factors, letting k be learnt from the data. Because the marginal distribution of X is vastly bigger than the conditional regression of the response given the predictors, the least eigenvalue scenario can still occur, even when letting k be inferred, if the priors over all of the parameters of the model are not carefully chosen.

First, the marginal likelihoods required to properly adjudicate between factor models of different sizes, whatever the prior used, are not feasible to obtain at present. Various approximations have been studied [Lopes and West, 2004], none of which speak directly to the concern of making sure a particular response is well explained. We address this explicitly in the following sections.

Second, the problem of finding a low-rank factor decomposition for a known covariance matrix is a computationally challenging problem [Fazel, 2002] with no known analytic solution. As a consequence, a closed form expression of our prior on k cannot be specified in terms of its most intuitive interpretation, the optimum value of

$$\begin{aligned} \text{minimize } & \text{rank}(\Sigma - \Psi) && (17) \\ \text{subject to } & \Sigma - \Psi \succeq 0 && (18) \end{aligned}$$

where $A \succeq 0$ denotes that A is positive semidefinite and Ψ is a diagonal covariance matrix. This problem, called the Frisch problem [Frisch, 1934], has a long pedigree in psychometrics, econometrics and control theory. It represents a significant barrier to Bayesian solutions to the model selection problem for factor models, because modern-scale problems where p is in the hundreds or thousands or greater, may not admit exact solution. In fact, the problem belongs to the computational complexity class NP-Hard [Vandenberghe and Boyd, 1996]. As a result, k must rather be interpreted in terms of arbitrary constraints placed on elements of the loadings matrix B which then interact with the prior for k in uncertain ways.

Our approach to chosing k is heuristic, setting $k \propto n$. With insufficient data to accurately recover more than n factors from a mere n observations, we shrink many of the p possible factor loadings to be identically zero. This can be thought of as an approximation to a prior under which all but k columns of B are centered at zero with sufficiently high precision that n observations will not appreciably move the posterior on these columns. The matter of most adequately explaining the variance of a particular response variable is then handled by modifying the likelihood for that response, conditional on this fixed k.

2 Partial factor regression

2.1 Specification

To address the least-eigenvalue situation, partial factor regression posits a lower dimensional covariance structure for the predictors, but permits the relationship between the predictors and the response to be linear in up to p dimensions. This is done by using the following covariance structure for the joint Normal distribution:

$$\begin{aligned} \begin{pmatrix} X \\ Y \end{pmatrix} &\sim \text{N}(0, \Sigma) \\ \Sigma &= \begin{bmatrix} BB^t + \Psi & V^t \\ V & \omega \end{bmatrix}. \end{aligned} \tag{19}$$

The difference between (27) and (15) is that here V is not required to equal θB^t. The matrix B is still a $p \times m$ matrix with $m \leq n \ll p$ so that the $p \times p$ predictor covariance matrix is constrained to the $BB^t + \Psi$ form, but, the full covariance matrix Σ is not simultaneously restricted. This way, the response can depend on directions in predictor space which are not dominant directions of variability, but inference and prediction still benefit from regularization of Σ_X.

Equally as critical, the prior on V may be conditioned on Σ_X. This hierarchical specification arises from the jointly Normal distribution between X, Y, and the m latent factors, which have covariance:

$$\text{cov}\begin{pmatrix} X \\ f \\ Y \end{pmatrix} = \begin{bmatrix} BB^t + \Psi & B^t & V^t \\ B & I_k & \theta^t \\ V & \theta & \omega \end{bmatrix}. \tag{20}$$

From this covariance, the conditional moments of the response can be expressed as:

$$\begin{aligned} E(Y \mid f, X) &= \theta f + \{(V - \theta B^t)\Psi^{-\frac{1}{2}}\}\{\Psi^{-\frac{1}{2}}(x - Bf)\} & (21) \\ \text{var}(Y \mid f, X) &= \omega - [V \quad \theta]\Sigma_{X,f}^{-1}[V \quad \theta]^t \quad \equiv \sigma^2. & (22) \end{aligned}$$

A natural prior for V, conditional on θ, B and Ψ might be

$$V \sim \text{N}(\theta B^t, w^{-1}\Psi),$$

implying that a priori the error piece plays no role in the regression. A reasonable choice of independent Normal prior on θ would be

$$\theta \sim \text{N}(0, I_k),$$

because the scale of the factors are set to have unit variance. All together, the model may be expressed as

$$\begin{aligned} X \mid B, f, \Psi &\sim \text{N}(Bf, \Psi) \\ Y \mid X, B, \theta, V, f, \Psi, \sigma^2 &\sim \text{N}(\theta f + \{(V - \theta B^t)\Psi^{-\frac{1}{2}}\}\{\Psi^{-\frac{1}{2}}(x - Bf)\}, \sigma^2) \\ V \mid \theta, B, \Psi &\sim \text{N}(\theta B^t, w^{-1}\Psi), \\ f &\sim \text{N}(0, I_k) \\ \theta &\sim \text{N}(0, qI_k) \\ b_{jg} \mid \psi_j, \tau_g &\sim \text{N}(0, \tau_g^{-1}\psi_j), \quad g = 1, \ldots, k, \quad j = 1, \ldots, p. \end{aligned}$$

The conditional regression parameters now borrow information from the marginal distribution via the prior – we have centered the regression at the pure factor model. However, the data may steer us away from this assumption.

Note that the marginal prior on V is given by

$$V \mid B, \Psi \sim \text{N}(0, BB^t + \Psi) = \text{N}(0, \Sigma_X).$$

Because $\beta = V\Sigma_X^{-1}$,

$$\begin{aligned} \text{cov}(\beta) &= \Sigma_X^{-1}\Sigma_X\Sigma_X^{-1} & (23) \\ &= \Sigma_X^{-1}, & (24) \end{aligned}$$

which mirrors the population version of Zellner's g-prior described above.

For computational considerations and ease of interpretation, we may work with a slightly reparametrized model, defining

$$\Lambda = (V - \theta B^t)\Psi^{-\frac{1}{2}} \tag{25}$$

and using the equivalent independent prior

$$\Lambda \sim \text{N}(0, w^{-1}I). \tag{26}$$

Note that $\Lambda = 0$ represents a pure factor model and that this prior is independent of the other parameters. The revised expression for our (latent) regression becomes

$$Y = \theta f + \Lambda\Psi^{-\frac{1}{2}}(X - Bf) + \epsilon, \qquad \epsilon \sim \text{N}(0, \sigma^2). \tag{27}$$

Notice that this formulation of the joint Normal model means that the conditional distribution of the response will be robust to decisions to fix k at a particular value or to the prior chosen if learning k is attempted. Similarly, the specific priors on V, B and θ are not so critical as long as $\mathrm{E}(V) = \theta B^t$. In particular, the prior on B may now benefit from sophisticated nonparametric specifications [Bhattacharya and Dunson, 2010], without any danger of affecting the conditional regression in unforeseen ways. Additionally, placing strong shrinkage priors on Λ could improve estimates of this high dimensional parameter [Carvalho et al., 2010]. The benefit of the partial factor decomposition stands apart from any of these individual modeling decisions.

By decoupling the predictor distribution from the conditional distribution, prior specification on the potentially ultra-high dimensional predictor space does not affect our lower dimensional regression in counterproductive ways. At the same time, the hierarchical prior on the regression parameters facilitates the borrowing of information that is necessary in the $p \gg n$ setting.

2.2 Comparison to ridge regression, partial least squares and factor models

For our simulation study, we let $p = 80$ and $n = 50$. Of the fifty observed values, only 35 observations are labeled with a corresponding Y value. For each of 150 such data sets, drawn at random, the remaining 15 unlabeled values were predicted using the posterior mean imputed value.

We compare four methods: partial factor regression (PFR), ridge regression (Ridge), partial least squares (PLS), and Bayesian factor regression (BFR).

The data was generated according to the following recipe.

1. Draw $k \sim \text{Uniform}(\{1, \cdots, n-1\})$.
2. Generate a matrix A of size $p \times k$ with independent standard Normal random variables.
3. Generate a $k \times k$ diagonal matrix D with elements drawn from a half-Cauchy distribution.
4. Set the true loadings matrix $B \equiv AD/\|AD\|$ where the $\|\cdot\|$ is the Frobenius norm.
5. The elements of Ψ are drawn independently as folded-t random variables with 5 degrees of freedom and scale parameter 0.1.
6. Lastly, θ was drawn by first drawing a folded-t scale parameter and then drawing a mean zero random variable with corresponding scale.

Finally, we consider two scenarios. In the first case, the elements of θ and D are reordered, so that the highest absolute value of D corresponds to the highest absolute value of θ, the second highest corresponds to the second highest, etc. This is a favorable case for the assumptions of ridge regression and factor models in that the response depends most on the directions of highest variability in predictor space. For the second case the elements of θ and D are arranged in reverse, so that the smallest absolute value of θ is associated with the largest absolute value of D. In this way the highly informative directions in predictor space are least informative of the response in terms of variation explained.

To compare the average behavior of these methods on a wide range of data we may look at the paired hold out error on each of the sets. We record the frequency that each method was the best performing method, the average percentage error relative to the best method, and also the average absolute error. The first measure records how often we should expect a method to be the best method to use on a randomly selected data set, so that higher numbers are better. The second column reflects how far off, on average, a given method performs relative to the best method for a given data set, so that smaller numbers are better. The final column gives the average error, so that numbers closer to zero are better.

We observe that in the favorable setting the pure factor model is quite often the best model of the four, as shown in the first column. However, we notice also that when it is not the best, it performs, on average, much worse than the best method, as shown in the second column. This is the impact of the bias. Next, we note that while ridge regression moderately outperforms the partial factor model in terms of overall mean squared error, we see that on average partial factor regression is closer to the best performing model. Relatedly, it is the partial factor model that is most often the best model.

In the unfavorable setting, things are more unambiguously in favor of the partial factor model. In this setting, as expected, the partial factor model outperforms ridge regression by all three measures. Again, the pure factor model is crippled by its strong bias. These findings are entirely consistent with those reported

Table 1: Favorable setting

	% time optimal	ave. % over optimal	scaled MSE
PFR	36	37	0·56
Ridge	19	48	0·53
PLS	16	62	0·63
BFR	29	727	1

Table 2: Unfavorable setting

	% time optimal	ave. % over optimal	scaled MSE
PFR	43	27	0·76
Ridge	17	45	0·79
PLS	10	64	0·89
BFR	30	387	1

in George and Oman [1996]. This strong performance profile recommends the partial factor approach as a sound default method for regression in the $p \gg n$ setting.

2.3 Chemometric Example

This trade-off between various linear regression methods appears in routine data analysis. Here we study an example from chemometrics, taken from Varmuza and Filzmoser [2009]. In this example, the response variable is a continuous outcome, the Lee gas chromatograph retention index. The observations are $n = 209$ distinct chemical compounds and $p = 467$ chemical descriptors are used as predictor variables. In this context, different linear regression techniques are treated as calibration methods. Hold out prediction error is used to adjudicate between the various techniques.

This situation is unsatisfactory from a statistical viewpoint, because the different linear regression methods vary purely in terms of their prior biases. Using cross validation to judge amongst them is simply uncovering which assumption is better for the data at hand, but in an ad hoc manner. The partial factor model builds in this uncertainty directly, favoring a low dimensional factor model, but permitting the data to dictate otherwise. As in our simulation study, we find that the partial factor model outperforms the other methods in terms of out of sample prediction, yielding a standard error of prediction of 8·1, compared to 9·6 for the next closest model (partial least squares with step-wise variable selection).

Two additional comments are in order. First, the originating textbook concludes that ridge regression is the "best model", but only after altering the criteria to use a trimmed prediction error, discarding especially bad predictions. Interestingly, such conspicuous misfires are likely a result of the least-eigenvalue scenario, where ridge regression is liable to over-regularize.

Second, the predictor variables in this example are most decidedly not actually drawn from a multivariate Normal distribution – indeed, some of the predictors are categorical and more than a few are bimodal. However, this misspecification does not harm the conditional regression or even the borrowing of information realized by the joint model and the hierarchical prior.

3 Sparse partial factor regression

3.1 Sparsity priors for variable selection

Consider the row vector Λ in (25). If $\lambda_j = 0$ predictor X_j appears in the regression of Y only via its dependence on the latent factors. Further, if we assume that θ is not identically zero so that Y has some relation to the latent factors, then we see that if $b_j = 0$ (so that dimension j does not load on any of the factors) and $\lambda_j = 0$, then $\beta_j = 0$ necessarily. That is, if X_j is not related to any of the latent factors governing the predictor covariance and additionally is not idiosyncratically correlated with Y via λ_j, then

X_j does not feature in our regression. The reverse need not hold; the net effect of X_j on Y can appear insignificant if X_j has a direct effect on the response, but is positively correlated with variables having the opposite effect.

Partial factor regression helps distinguish between these two scenarios, because the framework permits sparsity to be incorporated in each of three separate locations, with the following easy interpretations.

1. Does variable X_j load on latent factor $f_g \iff (b_{jg} = 0$ versus $b_{jg} \neq 0)$?
2. Does Y depend on the residual of element X_j; is X_j important for predicting Y above and beyond the impact of the latent factors $\iff (\lambda_j = 0$ versus $\lambda_j \neq 0)$?
3. Does Y depend on latent factor $f_g \iff (\theta_g = 0$ versus $\theta_g \neq 0)$?

This decomposition avoids the unsatisfactory choice of having to decide which of two variables should be in a model if they are very highly correlated with one another and associated with the response. Rather it allows one to consider the common effect of two such variables in the form of a latent factor, and then to consider separately if both or neither should enter into the model residually via the parameter Λ. Earlier work has keyed on to the idea that covariance regularization is useful for variable selection problems [Jeng and Daye, 2010], but the intuitive decomposition described above follows directly from the generative structure of the partial factor model.

Such a variable selection framework may be implemented with the usual variable selection point-priors on θ, Λ and B. Previous work incorporated such priors for the elements of B [Carvalho et al., 2008]. Alternatively, shrinkage priors may and thresholding may be used to achieve a similar effect [Carvalho et al., 2010].

3.2 Sufficient dimension estimation

In the case of multivariate Normal random variables, a factor decomposition of the covariance matrix, in combination with point mass priors as described above, admits a ready characterization of the sufficient subspace [Cook, 2007, Cook and Forzani, 2008] with respect to the response Y. A sufficient subspace is the span of a projection of the predictors which is sufficient to characterize the conditional distribution of the response.

In the factor model setting, we can calculate the dimension of this space as follows [Mao et al., 2010]. Let θ_Y denote the nonzero elements of θ in the partial factor parametrization and denote by B_Y the corresponding columns of B and likewise let B_X denote the remaining columns. Then, if $\Lambda = 0$, the conditional distribution of Y given X can be characterized purely in terms of

$$\begin{aligned} E(Y \mid X) &= \theta B^t (BB^t + \Psi)^{-1} X \\ &= \theta_Y B_T^t (B_Y B_Y^t + B_X B_x^t + \Psi)^{-1} X \\ &\equiv \theta_Y B_Y^t (B_Y B_Y^t + \Delta)^{-1} X \\ &= \theta_Y [\mathrm{I} - B_Y^t \Delta^{-1} B_Y (\mathrm{I} + B_Y^t \Delta^{-1} B_Y)^{-1)}] B_Y^t \Delta^{-1} X, \end{aligned} \tag{28}$$

where $\Delta \equiv B_X B_X^t + \Psi$, showing that X enters this distribution only via $B_X^t \Delta^{-1} X$. Thus, the rank of B_Y is the dimension of the sufficient subspace, as long as we have a pure factor model. We have already seen, however, that while a covariance matrix may be relatively well approximated by a small number of factors, these factors alone may not span the sufficient subspace, so that θ is estimated to be approximately zero and σ^2 is upward biased.

At the same time, in the $p \gg n$ setting we do not have the luxury of letting $k = p$; similarly learning k is a demonstrably hard problem, as has been mentioned. Accordingly, what we propose is to estimate $\Pr(\Lambda = 0 \mid X, Y)$, the posterior probability that the sufficient subspace is less than $k = \mathrm{rank}(B)$. Further, by monitoring the number of nonzero elements of θ in our sampling chain, we can further estimate the sufficient dimension, conditional on it being less than k.

This approach mirrors our approach to heuristically selecting k for regularization. The idea is that for a fixed amount of data, we only wish to entertain a restricted set of hypotheses – in this case, that the subspace is less than or equal to a prespecified k – but would simultaneously like to quantify the evidence that the truth lies outside this restricted class. The benefit here, as before, is that this reduced set of hypothesis – including the "none of the above" hypothesis – is more amenable to prior specification. And it is precisely

in the $p \gg n$ setting that prior specification is most relevant to posterior inference. Again, setting k as a function of n can be thought of as choosing the prior precision parameters in such a way as to manually enforce capacity control of model complexity.

4 Discussion

In the $p \gg n$ setting, inference and prediction can often be improved by making structural simplifications to the statistical model. In a Bayesian framework this can be accomplished by positing lower dimensional latent variables which govern the joint distribution between predictors and the response variable. An inherent downside to this approach is that it requires specifying a high dimensional joint sampling distribution and the associated priors. By simple virtue of the high dimensionality this task is difficult, particular with respect to appropriately modulating the implied degree of regularization of any given conditional regression.

The partial factor model addresses this difficulty by reparametrizing the joint sampling model using a compositional representation, allowing the conditional regression to be handled independently of the marginal predictor distribution. Specifically, this formulation of the joint distribution realizes borrowing of information via a hierarchical prior rather than through a fixed structure imposed upon the joint distribution.

Here we have examined the simplified setting of a joint Normal distribution. However, the idea of utilizing a compositional representation in conjunction with a hierarchical prior can be profitably extended to many joint distributions. In particular, one may specify the joint distribution directly, building in borrowing of information by design. For example, the form of the conditional moment for the partial factor model suggests the following nonlinear generalization:

$$E(Y \mid f, X) = g(f) + h(X - E(X \mid f)), \tag{29}$$

where perhaps g and h denote smooth functions to be inferred from the data. Here, the smoothness assumptions for g and h could be different; specifically the prior on h could be conditioned on properties of g. More generally, the partial factor model is a special case of models of the form:

$$\begin{aligned} f(Y, X \mid \Theta) &= f(X \mid \theta_X) f(Y \mid X, \theta_X, \theta_Y) && (30) \\ \pi(\Theta) &= \pi(\theta_Y \mid \theta_X)\pi(\theta_X). && (31) \end{aligned}$$

Such models alleviate the burden of having to get the high dimensional distribution just right in all of its many details. As such, it represents a robust method for fashioning data-driven prior distributions for regression models.

Appendix 1

Sampling Strategy

Here we outline our strategy for constructing a Markov chain with the partial factor posterior as its stationary distribution. This strategy is of general interest because it applies to any model expressible as in 30. The key assumption is that we are able to draw from $\pi(\theta_X \mid X)$, the posterior of the marginal model for the predictor distribution . In our case, this is possible using a Gibbs sampler with conjugate full conditionals, as described in Carvalho et al. [2008].

Our approach is to use a joint Metropolis step for parameters $\Theta \equiv \{\theta_X, \theta_Y\}$, using proposal distribution

$$\begin{aligned} q(\theta_X, \theta_Y) &= \{\pi(\theta_X \mid X)\}\{\pi(\theta_Y \mid X, Y, \theta_X)\} \\ &= \{c(X) f(X \mid \theta_X)\pi(\theta_X)\}\{(c(X, Y, \theta_X) f(Y \mid X, \theta_X, \theta_Y)\pi(\theta_Y \mid \theta_X)\}, \end{aligned} \tag{32}$$

where $c(\cdot)$ denotes the relevant normalizing constant. Note that this differs from the true posterior just in that the first factor is not conditioned on Y, as it ought to be. Because the target distribution may be expressed as

$$\pi(\theta_X, \theta_Y \mid X, Y) = c(X, Y) f(X \mid \theta_X) f(Y \mid X, \theta_X, \theta_Y)\pi(\theta_X)\pi(\theta_Y \mid \theta_X), \tag{33}$$

we obtain many canceling terms so that our acceptance ratio takes the form

$$\begin{aligned}\alpha(\tilde{\Theta}, \Theta) &= \left\{ \frac{\pi(\tilde{\theta}_X, \tilde{\theta}_Y \mid X, Y)}{\pi(\theta_X, \theta_Y \mid X, Y)} \right\} \left\{ \frac{q(\theta_X, \theta_Y)}{q(\tilde{\theta}_X, \tilde{\theta}_Y)} \right\} \\ &= \frac{c(X, Y, \theta_X)}{c(X, Y, \tilde{\theta}_X)}.\end{aligned} \tag{34}$$

Disregarding some variance terms for simplicity, the partial factor model takes $\theta_Y = \{\Lambda, \theta, \sigma^2\}$ and $\theta_X = \{B, \Psi, F\}$ so that

$$c(X, Y, \theta_X) = |S|^{-\frac{1}{2}}, \tag{35}$$

where S denotes the posterior covariance matrix for the parameters of the conditional Normal regression given in (27).

A Acknowledgment

P. R. Hahn and C.M. Carvalho would like to acknowledge the support of The University of Chicago Booth School of Business. C.M. Carvalho also acknowledges the support of The Harrington Foundation at The University of Texas. P.R. Hahn and S. Mukherjee would like to acknowledge the support of NIH Grant P50 GM 081883 and S. Mukherjee would also like to acknowledge the support of NSF Grant DMS-07-32260 and NIH Grant R01 CA123175-01A1.

References

O. Aguilar and M. West. Bayesian dynamic factor models and variance matrix discounting for portfolio allocation. *Journal of Business and Economic Statistics*, 18:338–357, 2000.

A. Bhattacharya and D. B. Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 2010.

C. M. Carvalho, J. Lucas, Q. Wang, J. Nevins, and M. West. High-dimensional sparse factor modelling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 2008. to appear.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97 (2):465–480, 2010.

R. D. Cook. Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1):1–26, 2007.

R. D. Cook and L. Forzani. Principal fitted components for dimension reduction in regression. *Statistical Science*, 23(4):485–501, 2008.

D. Cox. Notes on some aspects of regression analysis. *Journal of the Royal Statistical Society Series A*, 131:265–279, 1968.

M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, March 2002.

L. Forzani. Principal component analysis: A conditional point of view. December 2006.

R. Frisch. Statistical confluence analysis by means of complete regression systems. Technical Report 5, University of Oslo, Economic Institute, 1934.

S. Fruhwirth-Schnatter and H. Lopes. Parsimonious bayesian factor analysis when the number of factors is unknown. Technical report, University of Chicago Booth School of Business, 2009.

E. I. George and S. D. Oman. Multiple-shrinkage principal component regression. *Journal of the Royal Statistical Society Series D (The Statistician)*, 45(1):111–124, 1996.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, 2001.

H. Hotelling. The relationship of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, 10:69–79, 1957.

X. J. Jeng and Z. J. Daye. Sparse covariance thresholding for high-dimensional variable selection. *Statistica Sinica*, 2010.

I. T. Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society, Series C*, 31(3):300–303, 1982.

F. Liang, S. Mukherjee, and M. West. The use of unlabeled data in predictive modeling. *Statistical Science*, 22(2):189–205, 2007.

F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger. Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103:410–423, March 2008. URL `http://ideas.repec.org/a/bes/jnlasa/v103y2008mmarchp410-423.html`.

H. Lopes. Factor models: An annotated bibliography. *Bulletin of the International Society for Bayesian Analysis*, 2003.

H. Lopes and M. West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14:41–67, 2004.

K. Mao, F. Liang, and S. Mukherjee. Supervised dimension reduction using Bayesian mixture modeling. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010.

L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM*, 38(1):49–95, March 1996.

K. Varmuza and P. Filzmoser. *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press, 2009.

M. West. Bayesian factor regression models in the "large p, small n" paradigm. In J. M. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, editors, *Bayesian Statistics 7*, pages 723–732. Oxford University Press, 2003.

A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. Elsevier, 1986.